Is Gene Duplication a Viable Explanation for the Origination of Biological Information and Complexity?

All life depends on the biological information encoded in DNA with which to synthesize and regulate various peptide sequences required by an organism's cells. Hence, an evolutionary model accounting for the diversity of life needs to demonstrate how novel exonic regions that code for distinctly different functions can emerge. Natural selection tends to conserve the basic functionality, sequence, and size of genes and, although beneficial and adaptive changes are possible, these serve only to improve or adjust the existing type. However, gene duplication allows for a respite in selection and so can provide a molecular substrate for the development of biochemical innovation. Reference is made here to several well-known examples of gene duplication, and the major means of resulting evolutionary divergence, to examine the plausibility of this assumption. The totality of the evidence reveals that, although duplication can and does facilitate important adaptations by tinkering with existing compounds, molecular evolution is nonetheless constrained in each and every case. Therefore, although the process of gene duplication and subsequent random mutation has certainly contributed to the size and diversity of the genome, it is alone insufficient in explaining the origination of the highly complex information pertinent to the essential functioning of living organisms. © *2010 Wiley Periodicals, Inc. Complexity 16: 17–31, 2011*

Key Words: gene duplication; biological complexity; evolutionary divergence; compensatory mutation; conservation of information

JOSEPH ESFANDIAR HANNON BOZORGMEHR

Joseph Esfandiar Hannon Bozorgmehr is Manchester M9 4GQ, United Kingdom (e-mail: bozorgmehr@hotmail.co.uk)

1. INTRODUCTION

1.1. The Efficacy of Natural Selection

One of the singular issues in molecular biology and evolution concerns the origins of the distinct exonic sequences and motifs that contribute to the functionality of the genome and to organismic complexity. Indeed, the cause of such a huge proliferation of genetic information, coding for polypeptides as small as 49-residue echistatin to those such as titin, a gigantic protein found in muscle tissue and consisting of over 30,000 amino acids remains an elusive and unsolved problem in the study of biological origins. It is presumed that the genes of all

 Vol. 16, No. 6
DOI 10.1002/cplx.20365
Published online 22 December 2010 in Wiley Online Library (wileyonlinelibrary.com)

extant and extinct species have evolved from a life-form with a protogenome [1]. Natural selection per se is a poor candidate to explain such an evolution of complexity [2], as it is disposed to conserve the existing structure and organization of genes, and their essential information content, resulting in functional stasis [3].

Research into the evolution of genes has shown that the peptides they code for are of a finicky and precarious nature, both marginally stable and prone to aggregation [4]. Protein folding happens to be a highly complex and synergistic process, involving a number of epistatic relationships among many residues. This phenomenon, compounded with the issue of interactions between protein molecules, can significantly complicate adaptive evolution such that in the majority of cases the overall effects on reproductive fitness are very slight [5, 6]. Many arguably "beneficial" mutations have been observed to incur some sort of cost and so can be classified as a form of antagonistic pleiotropy [7].

Indeed, the place and extent of natural selection as a force for change in molecular biology have been questioned in recent years [8]. Detecting the incidence of any beneficial substitutions in genes has so far relied on statistical inferences as empirical evidence is less readily available. In many instances, nonsynonymous changes and shifts in allelic diversity may be induced by factors that can serve to imitate selective effects—biased gene conversion, mutational and recombinational hotspots, hitchhiking, or even neutral drift being among them [9]. Moreover, several well-known factors such as the linkage and the multilocus nature of important phenotypes tend to restrain the power of Darwinian evolution, and so represent natural limits to biological change [10]. Selection, being an essentially negative filter, tends to act against variation including mutations previously believed to be innocuous [11]. For example, *PABPC1* is a polyadenylate-binding protein used in translation initiation in both humans and mice [12]. Although there are 92 nucleotide differences in the translated region of the respective orthologous genes, these are all synonymous except in just two codons where Asp has been replaced at residue 209 with Glu and Thr with Ser at residue 576—both similar amino acids. However, it is also clear that the gene's role is essential and that any functional divergence in this particular case is unnecessary.

1.2. Duplication as a Potential Driving Force Behind Molecular Evolution

Gene duplication offers the prospect of a respite from stringent purifying/negative selection [13]. This is because only one gene locus needs to be functional, meaning that any paralog is freer to diverge allowing for changes, promoted by near neutral drift, which would not normally be tolerated in the case of a singleton. It is thought that suboptimal and deleterious changes may become fixed and accumulate through a more permissive regime of selection [14], such that they fortuitously combine to produce a novel adaptive function. However, any evolutionary development must be tempered by the impact of any changes on protein structure and stability [15] and not just the peptide sequence itself.

Although it may be inefficient and costly for the cell to produce identical surplus proteins, and which can lead to cases of unwanted overexpression and harmful phenotypes [16], this can also prove to be beneficial by providing a useful double dosage [17]. Similarly, the role of duplicate genes in facilitating alternative metabolic pathways and regulatory interactions [18] is another important factor.

Duplication, including instances of intragenic amplification, can occur by way of unequal chromosomal crossovers, the retropositioning of spliced mRNA, and copying of a whole chromosome or even an entire genome—the persistence of entire gene networks helps to explain the presence of polyploidy in plants [19].

However, genomic studies have revealed that active duplicates may nonetheless be selected for their redundant utility [20], as they can serve as backups when a mutation inflicts damage to a sister site [21, 22]. This means that any changes made to them are liable to be selected against if they impair this masking ability and its contribution to genomic robustness. This may explain, in part, the huge effect of duplicates in shaping both prokaryotic and eukaryotic genomes, and their evolutionary preservation [23]. Another phenomenon involved in the retention of duplicate genes is "subfunctionalization," namely the differential partitioning of function or expression [24]. Here, redundant functions will degenerate at random from the daughter copies until their joint function matches that of the parent gene [25].

Were selection to be completely relaxed and any manner of changes permitted, this would only serve to guarantee complete degeneration. It would invariably lead to the introduction of null and nonsense mutations, scrambling the open reading frame (ORF), and degrading the cisregulatory elements involved in transcription—leading to the gene's pseudogenization. Thus, a measure of purifying/stabilizing selection seems necessary for duplicate preservation, and any evolutionary divergence would proceed under a relaxed regime rather than none at all.

Moreover, in terms of population size, Kimura's diffusion approximation [26] makes it abundantly clear that in diploid populations of a normal size, typically for those of $N > 10{,}000$, even the slightest degree of negative selection is sufficient to prevent any deleterious allele from surviving and increasing in frequency to the point of fixation or near fixation. This would mean that any major changes in

DOI 10.1002/cplx

both gene singletons and duplicates alike would tend to occur in smaller populations, where drift is much stronger and selection is weaker.

2. THE EVOLUTION OF GENETIC INFORMATION

The purpose of this study is to determine the existence and extent of any novel information produced as a consequence of gene duplication. At stake is whether there is sufficient supporting evidence that the digitally communicated instructions [27] encoded in DNA could have been constructed through known evolutionary processes, or whether the data suggests that an alternative explanation is required as in all codified nonbiological information. Therefore, this would serve as a means of assessing the current arguments regarding the origins of biological and genomic complexity.

2.1. The Information Conundrum

Although the nucleotide sequences in DNA are commonly understood to carry/convey biological "information" [28], a precise scientific delineation for the term in the context of genetics is often found to be lacking. Therefore, it is impossible to test any hypothesis regarding the creation of new genetic information without offering at least a conceptual definition of what information means and what the criterion is for identifying it. In Shannon's theory [29] of communication, information is termed the "reduction in uncertainty," where entropy is the measure of any stochastic dependencies—the greater the level of uncertainty that exists in a particular situation, the less likely it is to predict the behaviors and outcomes because of the presence of random noise. Therefore, information is that which denotes a degree of determinism in a known relationship, although this would also have to involve a large measure of contingency to permit as many possible combinations to be conveyed. In the framework of molecular biology, information would refer to the inherent functionality of gene products: i.e., how they interact with the biochemical environment in which they operate.

Therefore, I have decided to define any gain in exonic information as: "The qualitative increase in operational capability and functional specificity with no resultant uncertainty of outcome." The two parts of the statement are complementary, because an appreciably great degree of specificity is required to reduce any uncertainty and problems regarding behavior and effect: this is especially true in the case of enzymes that catalyze only particular reactions, and to the exclusion of all others. A random mutation in the active site could well lead to an "advantageous" outcome in a particular environment owing to a shift in catalytic activity. However, the evidence suggests this would entail an alteration in the particular specificity pattern [30]. Therefore, it would mean that an increase of uncertainty and more erratic behavior, with respect to the overall and net effect(s), is a consequence of such a development.

2.2. The Relationship Between Sequence, Function, and Evolutionary Divergence

Usually, it is safe to say that homologs share basically the same function and that many changes in sequence are not consequential. However, this is very much a general rule. A single amino acid replacement in a carboxyl esterase in blowflies confers organophosphorus insecticide resistance [31], although this is because of a loss in the primary enzymatic activity. Many synonymous changes have indeed been identified with codon usage bias, contributing to splicing and translational efficiency [32]. A study has found that there exists a threshold at $\sim$ 50% sequence similarity below which functional divergence is enhanced [33]. Orthologs performing the same function should be under the same selective constraints and evolve at the same rate. But in the case of paralogs, there is a relaxation of purifying selection, and distinguishing loss of constraint from rapid evolution driven by adaptation is difficult because the loss of constraint often precedes any potential neofunctionalization [34].

2.3. Testing for the Role of Natural Selection in the Creation of Novel Functionality

Detecting the effect of Darwinian positive selection—whereby an allele is supposed to increase in frequency because it confers a reproductive advantage—is not an exact science by any means, and it relies on statistical-based inferences that leave much to interpretation. Even if adaptive mutations have been prominent in a gene, it is not accurate to necessarily infer that any new functionality has arisen. All it means is that an allele has contributed to a gain in reproductive fitness, and nothing beyond that. In many instances, as with the example above, a loss of function and regulation in a harsh or unusual environment can have a beneficial outcome and thus be selected for—bacteria tend to evolve resistance to antibiotics in such a way through mutations that would otherwise adversely affect membrane permeability [35]. The magnification of the importance of one or more loci is tantamount to artificial selection, but occurs in some cases during drastic environmental catastrophes, where a single trait might make a difference between survival or not.

Population genetics methods typically involve measuring levels of heterogeneity and polymorphism at sites including and in proximity to the one under investigation [36]. It can lead to confusing results because the effects of Darwinian selection are often the same as those of background selection—the purging of neutral alleles due to their spatial proximity to deleterious ones

DOI 10.1002/cplx

[37]: the case of the gene implicated in microcephaly likely being a controversial example of this [38]. Sequence alignment methods are preferred, especially where data from a sample of a population is not available. As such, three ratios were determined and used throughout to detect the probability of functional change [39].

i. The ratio of nonsynonymous to synonymous substitutions, dN/dS (ω), is regarded as the most obvious indication of adaptive change and functional shift [40]. In the case of neutral evolution, it would be around 1:1, but the proportion is skewed in favor of the former if positive selection is prevalent, whereas purifying selection is inferred when this is reversed. When comparing singletons in different phylogenetic lineages, this is a very powerful method, but in the case of duplicates more caution is required. As has been previously mentioned, there is an appreciably relaxed regime of selection in paralogous genes because only one need maintain the original function(s). As such, the rate of nonsynonymous substitutions may be much higher, not on account of adaptive evolution, but because purifying selection is far less stringent than it is for singletons.
ii. The transition to transversion ratio, ts/tv (κ), is also a useful test. Although there are twice as many possible transversions as there are transitions, the molecular mechanisms by which they are generated means that transitions (e.g., purine to purine) are more frequent than transversions (e.g., purine to pyrimidine). Notwithstanding mutational bias, the ratio can be seen as evidence for adaptation if the transversions greatly exceed transitions [41].
iii. The ratio of radical to conservative replacements, K_R/K_C, is a measure of the nature of the evolutionary changes in peptides. As many amino acids are chemically similar, they may also be relatively interchangeable—as with Val, Ile, and Leu—and so can be regarded as essentially neutral substitutions. Therefore, dN/dS may not reflect the significance of any divergence. If K_R/K_C is >1, then this could be suggestive of the fixation of beneficial mutations. However, such is the nature of context specificity within protein domains that a suboptimal but still conservative replacement at one site could require a compensatory [42] and more radical change at another. Although widely used, the method has been criticized for being too simple and shows nothing about actual changes in the behavior of the protein [43].

2.4. Aims of Investigation and Materials Used

Several familiar and exemplary cases of evolution following an initial gene duplication were chosen and categorized according to known mechanisms of divergence that include fusion, frameshift mutations, retroposition, internal amplification, and de novo recruitment. There is, of course, considerable overlap between these various mechanisms, although the primary focus is different for each case. The scope and remit of the investigation was limited to exonic sequences within the translated regions, thus largely avoiding regulatory areas and introns, where retrotransposon insertions are believed to be significant [44]. Although gene regulation and expression are important, it is the regions that code for protein sequences that comprise by far the primary source of biological information. All pertinent sequence data, both nucleotide and amino acid, were downloaded from the NCBI database and taken from where it is cited in the relevant literature. Standard alignment techniques for analyzing and illustrating the data were done using BLAST, with more advanced pair-wise ones using the ClustalW2 algorithm together with Emboss.

3. ANALYSIS OF GENE DUPLICATION BY EXAMPLE

3.1. Duplication and Gene Fusion: The Case of Sdic

Sdic is believed to be a flagellar dynein gene found only in *Drosophila melanogaster*—an example of a tandem duplicated chimeric gene "caught in the act" of evolving [45]. It was formed when two adjacent genes, *AnnX* (coding for a cell adhesion protein) and *Cdic* (encoding a cytoplasmic intermediate chain dynein), were first duplicated and one pair subsequently underwent a deletion-mediated fusion. *Sdic* is found to be composed of four paralogs having itself been duplicated twice over. The 5' untranslated region (UTR) and part of the promoter sequence of the gene derives from *AnnX*, whereas the translated part and all 300 base pairs (bp) of the 3′ UTR come from the *Cdic* gene. A sequence comparison of *Sdic2* and *Cdic* reveals that 522 out of 527 residues (99%) can be aligned without difficulty. *Sdic* has been observed to be expressed in the testes and incorporated into the sperm tail and this is because it has acquired a testis-specific core element, homologous with those of other promoter sequences, from the 3′UTR of *AnnX* [46]. It is unclear whether the element is a translational enhancer or has some other regulatory role in the *AnnX* gene such as, for example, in mRNA localization. Either way, the gene would seem to contribute to greater fecundity.

But, it is the loss of over 100 codons from *Cdic*'s N-terminus [47], involving at least two domains, that deprive the *Sdic* protein of the motifs necessary to enable it to interact with dynactin (a basic characteristic of cytoplasmic dyneins) and which represent the prin-

DOI 10.1002/cplx

cipal functional shift. Thus, *Sdic* is axonemal almost by default owing to the mass deletion of exonic information pertinent to cytoplasmic-specific operation (Figure 1). The gene's promoter has simply acquired features from preexisting coding sequences and information present in *AnnX*, whereas its translated region is virtually identical with the corresponding part of *Cdic*.

The distal and proximal conserved elements are also found to be very similar to those of the *Cdic* promoter. In addition, the 16 codons present at the N-terminus of *Sdic*, recruited from Cdic's third intron along with an 11 bp insertion, bear a tenuous resemblance to the amino ends of axonemal intermediate chain dyneins such as those for *oda6* and *AclC3* [48]. It is reasonable to assume that this small amount of exonization, allowing a previously noncoding region UTR to become the start site and initial part of the *Sdic* gene, is adaptive. As such, this could be interpreted as evidence for the de novo creation of novel information.

Further evidence for the role of selection in the development of *Sdic* includes a possible sweep found in the low levels of polymorphism across neighboring loci and a skewed frequency distribution of allelic variation. However, it is noted that a reduced level of heterozygosity in a region of low recombination, such as at the base of the X-chromosome where *Sdic* is located, is also consistent with background selection because of the effect of deleterious mutations [49]. Both analyses could in fact be correct. Although the number of nonsynonymous differences is greater than synonymous ones, as would be expected in a basic test for adaptive evolution, this is due to a bulk deletion and resultant frameshift occurring in the fourth domain (inherited from *Cdic*) that produced a string containing at least five novel characters. As this domain is believed to be nonfunctional in *Sdic*, it is more logical to infer the existence of a relaxed regime and decrease in selective constraints, than to assume any adaptive change. Therefore, the initial loss of information at the N-terminus because of relaxed selection was then compensated for by the recruitment of sequences from an intron of *Cdic* and the exons of *AnnX*. In this way, a nonfunctional cytoplasmic dynein "evolved" into an axonemal one through a process of copy, cut, and paste.

Divergence between the *Sdic* paralogs themselves has been very limited such that the translated regions of *Sdic2* and *Sdic4* actually share a 100% nucleotide identity and are functionally redundant. Although the gene is considered to be young, and formed within the last 2–3 million years, the short generational span of the fruit fly ($\sim$ 2 weeks) means that the evolutionary timescale may actually be rather long ($\sim$ 50 m generations).

It is possible that *Sdic* contributed to speciation and the emergence of the *melanogaster* line [50]. The most likely scenario involves a population bottleneck, migration, or founder effect [51]. Any reduction in effective population size would also produce a further relaxation of selective constraints as (nearly) neutral drift would predominate.

It appears that deletion in this instance was one of the necessary factors involved in gene fusion. As such, *Sdic* is shorter than *Cdic*, and this is true also for the hominoid oncogene, *TRE2*, which is 200 residues less than one of its parents, *USP32* [52]. This presents a problem in terms of explaining any accretion of cistron size with reference to the most naturally applicable evolutionary process. Deletion-mediated fusion also means that usually one of the genes is far less preserved than the other but in the case of *Kua-UEV,* however, the effect is additive because it has retained the original and separate functions of both its parents [53]. Although it may behave slightly differently, particularly with respect to intracellular localization, the information content has not appreciably changed.

3.2. Duplication and Frameshift Mutation

Already briefly mentioned in the previous section, another potential means by which new genes, with new exonic information, might arise is by way of a frameshift resulting in an entirely different ORF and peptide sequence. A case of just such a development was proposed by Ohno [54] in the case of a nylon oligomer hydrolase found in bacteria near sites involved in the production of the synthetic material. However, a study by Negoro et al. [55] found that the likely source was actually an esterase containing a β-lactamase fold. Two amino acid replacements in the catalytic cleft greatly increased the Ald-hydrolytic activity, in some measure already provided by a serine active site, necessary for the degradation of the oligomers. However, this does appear to have come at some cost to part of the esterolytic function and the enzyme does not have nearly the specificity constant and efficiency, with respect to its alternative functionality, of a hydrolase such as aminoacylase [56]. Therefore, although there is an appreciable gain in operational capability, no new information was generated that specified oligomer degradation.

Scherer and coworkers [57], using a search on BLAST, found that as many as 470 duplicated genes in humans had been affected by frameshift translation. However, frameshifts induced either by indels or by transposons (mobile elements) are themselves poor candidates for the generation of novel information because they almost inevitably incur premature stop codons [58], leading to protein truncation, in addition to scrambling part of the original reading frame. This is indeed evident in some of the genes presented in their study. *HTR3D* is a hydroxytryptamine (serotonin) receptor in humans, which is essentially the carboxyl terminus remnant of *HTR3C*. However, owing to the inherent modularity of a gene, the truncated daughter copy has retained at least part of the parental func-

DOI 10.1002/cplx

FIGURE 1

```
CDIC   MDRKAELERKKAKLAALREEKDRRRREKEIKDMEEAAGRIGGGAGIDKDQRKDLDEMLSSLGVAPVSEVLSSLSSVNSMTSDNSNTQTPDASLQATVNGQ   100
SDIC   ----------------------------------------------------------------------------MGLVLIKFLRSTYSTL   016
CDIC   SGGKKQPLNLSVYNVQATNIPPKETLVYTKQTQTTSTGGGNGDAHATDYYDEYNLNPGLEWEDEFTGDDEESSLQNLGNGFTSKLPPGYLTHGLPTVKDV   200
SDIC   SGGKKQPLNLSVYNVQATNIPPKETLVYTKQTQTTSTGGGNGD--------------VL-AFDAQ-GDDEESSLQNLGNGFTSKLPPGYLTHGLPTVKDV   100
CDIC   APAITPLEIKKETEVKKEVNELSEEQKQMIILSENFQRFVVRAGRVIERALSENVDIYTDYIGGGDSEEANDERSHARLSLNRVFYDERWSKNRCITSMD   300
SDIC   APAITPLEIKKETEVKKEVNELSEEQKQMIILSENFQRFVVRAGRVIERALSENVDIYTDYIGGGDSEEANDERSHARLSLNRVFYDERWSKNRCITSMD   200
CDIC   WSTHFPELVVGSYHNNEESPNEPDGVVMVWNTKFKKSTPEDVFHCQSAVMSTCFAKFNPNLILGGTYSGQIVLWDNRVQKRTPIQRTPLSAAAHTHPVYC   400
SDIC   WSTHFPELVVGSYHNNEESPNEPDGVVMVWNTKFKKSTPEDVFHCQSAVMSTCFAKFNPNLILGGTYSGQIVLWDNRVQKRTPIQRTPLSAAAHTHPVYC   300
CDIC   LQMVGTQNAHNVISISSDGKLCSWSLDMLSQPQDTLELQQRQSKAIAITSMAFPANEINSLVMGSEDGYVYSASRHGLRSGVNEVYERHLGPITGISTHY   500
SDIC   LQMVGTQNAHNVISISSDGKLCSWSLDMLSQPQDTLELQQRQSKAIAITSMAFPANEINSVVMGSEDGYVYSASRHGLRSGVNEVYERHLGPITGISTHY   400
CDIC   NQLSPDFGHLFLTSSIDWTIKLWSLKDTKPLYSFEDNSDYVMDVAWSPVHPALFAAVDGSGRLDLWNLNQDTEVPTASIVVAGAPALNRVSWTPSGLHVC   600
SDIC   NQLSPDFGHLFLTSSIDWTIKLWSLKDTKPLYSFEDNSDYVMDVAWSPVHPALFAAVDGSGRLDLWNLNQDTEVPTASIVVAGAPALNRVSWTPSGLHVC   500
CDIC   IGDEAGKLYVYDVAENLAQPSRDEWSRFNTHLSEIKMNQSDEV                                                            643
SDIC   IGDEAGKLYVYDVAENLAQPSRDEWSRFNTHLSEIKMNQSDEV                                                            543
```

The alignment of Cdic and Sdic (2 and 4) reveals the virtual identity of the corresponding coding regions in the genes. The N-terminus of Cdic, consisting of 100 codons, is missing in Sdic, and this means that Sdic lacks the motifs necessary with which to interact with dynactin. An intronic recruitment at the amino end has led to the exonization of 16 codons, whereas another deletion downstream, this time involving the loss of 16 codons, is present within the fourth domain from the 5′ end. This development has resulted in a frameshift that provided five novel characters in the sequence.

tionality, whereas the rest has been essentially ignored by purifying selection. Protein truncation in duplicates can also occur by way of a nonsense mutation resulting in a premature stop in translation: the G-type cyclin *CCNG1*, involved in the regulation of cell cycle kinases, is found to be missing an important "PEST" sequence at the C-terminus that is present in its paralog, *CCNG2* [59].

The authors cite as one such example of a possible frameshift the gene *SLC25A37*, a member of the mitochondrial solute carrier family. Indeed, an analysis reveals that *SLC25A37* was created, as shown in Figure 2, as a result of a bulk deletion together with a single nucleotide insertion in a copy of the likely parent, *SLC25A28*—although the exact sequence of events cannot be determined. As a result of the frameshift, 54 novel characters were generated but 22 were also deleted, casting doubt on the biochemical importance of this resulting minisequence. This would suggest that despite the extent of nonsynonymous differences evident in *SLC25A37*, these are likely to have been the result of relaxed purifying selection rather than any beneficial increase in information.

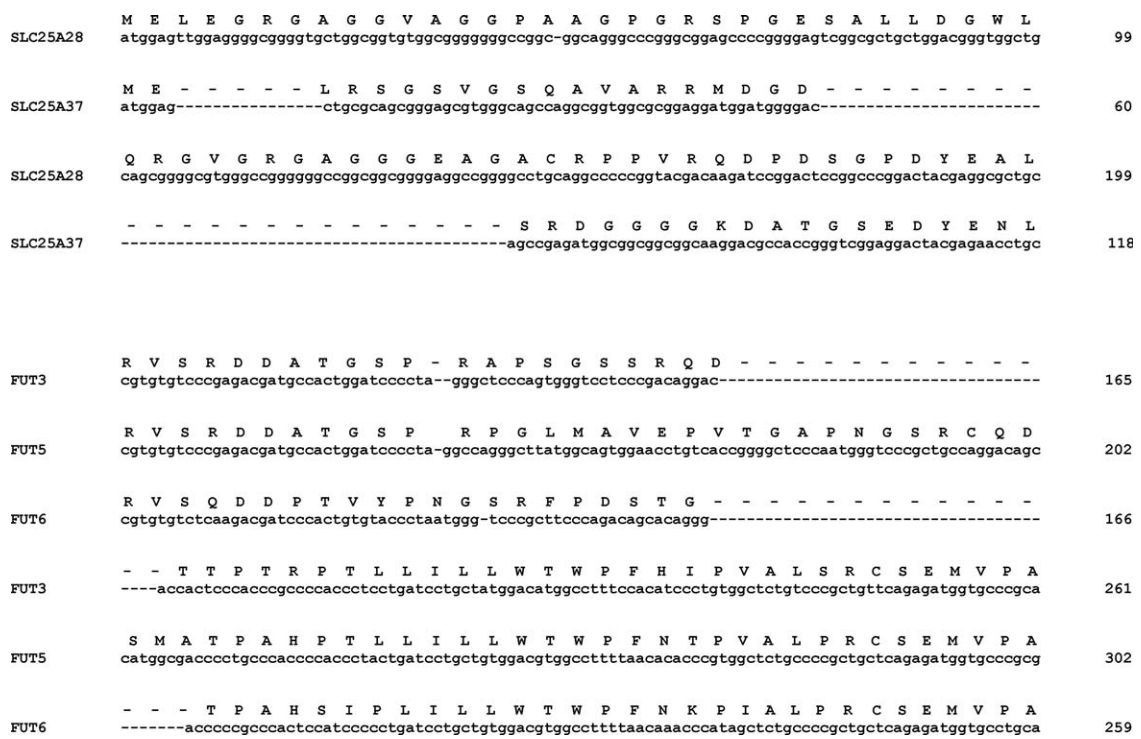
It would be useful to test for the effect of natural selection in the 302 codons of the gene downstream of the frameshift and where the original reading frame has been restored. Accordingly, it was observed that 213 aligned residues were identical and that the ratio (ω) of nonsynonymous to synonymous base pair substitutions was greater than 1.0 (169:114). The ratio (κ) of transitions to transversions was roughly equal (135:148), as was the ratio of radical to conservative amino acid substitutions (49:40). So, this would likely suggest that this reflects an overall structural realignment possibly to offset the radical changes and deletions at the N-terminus, rather than one representing any major functional shift.

Evolutionary divergence by frameshift mutation, and several other mechanisms, has also taken place in the *FUT* gene family in humans [60]. All but one of the nine genes are monoexonic and all code for the enzyme—fucosyltransferase—that transfers fucose on the terminal residues of glycans, albeit on a different variety of substrates. *FUT3* and *FUT6* are believed to be the most expressed members within the family and share a >90% nucleotide identity, displaying no discernibly significant functional differences. Both have diverged from a common ancestor, quite possibly *FUT5* itself, by way of a 40-bp deletion and resultant frameshift at the N-terminus—in much the same manner as the previous example. This is consistent with an inference for the relaxation of selective constraints and partial degeneration followed by a suppressing mutation.

Therefore, although frameshifts have the potential to cause more rapid sequence divergence than can individual point mutations, it is wrong to assume that they can produce any novel information even if they do

DOI 10.1002/cplx

FIGURE 2



Divergence by way of a frameshifting event in SLC and FUT genes in Homo sapiens. The regions within each gene sequence affected by indels are shown above. In the case of the mitochondrial solute carrier gene, SLC25A28, a 16-nt deletion at the N-terminus has occurred in a duplicated copy of it. This alone would have truncated the gene into two separate reading frames: 1-252 and 252-1079. However the insertion of adenine at nt position 48 suppresses any gene fission and restores the length of the original reading frame, giving rise to SLC25A37. In the FUT genes, a combination of deletions and at least one insertion in FUT3 and FUT6 caused a significant divergence in sequence from a common ancestor whose translated region would have resembled FUT5. In both cases, the reading frame is altered for a short region, involving the loss of many codons, before being reconstituted donwstream and thus demonstrating a conservation of information.

result in the emergence of novel characters within proteins. Therefore, a divergence in sequence need not result in a change in functionality or affect behavior, as the same information can be constructed using a number of different amino acid arrangements. In duplicates, and also singletons, changes may be compensatory and in response to prior degeneration rather than representing any innovation.

3.3. Gene Duplication and Retroposition: The Case of Jingwei and Adh

Another gene of interest to researchers of molecular evolution, found in *Drosophila yakuba* and *D. tessieri*, is *Jingwei* (*jgw*). Like *Sdic*, it is a chimeric gene except that it was formed from the retropositioning (by reverse transcription) of one gene into the duplicated copy of another [61]. This constitutes a type of ectopic recombination, otherwise known as exon shuffling. The first three exons are considered to be derived from a duplicated copy (*ynd*) of a gene that is expressed uniquely in the testes (*ymp*). Therefore, the N-terminal domain of *ynd* has donated the non-*Adh* portion of *jgw* and this appears to be well preserved by purifying selection, indicative of the retention of functionality and also of the modular structure and organization of the gene [62].

Adh is an alcohol dehydrogenase that occurs in many organisms and facilitates the interconversion between alcohols and aldehydes. The retrosequence of the gene was copied and inserted into the third intron of *ynd* and nine downstream exons of became pseudoexons, because transcription stopped at the terminating signal encoded in the *Adh*-derived exon. Initially, this led researchers to believe that *Jingwei* was nothing other than a pseudogene, and its exact function is still unknown. As such, the first 68 codons of the translated region are derived from the *ymp*/*ynd* gene, whereas the remaining 255 are derived from the original 272 codons of the

DOI 10.1002/cplx

translated region in the ancestral *Adh* gene. Betrán [63] and others speculate that the number of nonsynonymous changes should be regarded as evidence for rapid adaptive evolution. Indeed, only 92 of the original 272 residues remain (almost the minimum proportion to identify a homology), whereas the ratio (ω) of nonsynonymous to synonymous changes is astonishing—(332:58). The ratio (κ) of transitions to transversions (154:236) and radical to conservative amino acid replacements (113:49) is an indicative too of a substantial functional shift. But is that really what has happened? Is there another explanation to account for this?

Clearly, selective constraints have been relaxed as nine of the exons of the y*nd* gene were silenced by the initial act of retroposition, whereas the C-terminus of the intronless *Adh* retrosequence itself has been truncated by a frameshift with the resultant loss of 15 codons, for which a single nucleotide insertion would appear to be responsible. This loss of information is to be expected in a model of relaxed selection. The distribution of nonsynonymous changes in the *Adh* part of *Jingwei* is also found to be relatively uniform and not clustered in one particular region—the active site of *Adh* is indeed well preserved. This is either suggestive of widespread directional selection or, rather, random degeneration and destabilization: i.e., a failure to preserve the integrity of the sequence. However, the actual situation is likely to be more nuanced than either scenario would suggest. The introduction of deleterious changes could also set off a process and chain reaction of ensuing compensatory mutations observed in other genes [64, 65]—a proclivity toward physical stability being inherent in the nature of all proteins. Compensatory mutations would thus make up for any suboptimal or potentially damaging amino acid replacements elsewhere in the sequence, as opposed to back mutations that simply restore the ancestral residue. In this way, evolutionary divergence need not result in any change in the information content and functionality even if the resultant peptide sequence is substantially altered. Moreover, the effect of compensation following partial degeneration would be indistinguishable from any functional innovation because both are beneficial.

In vitro experiments, using a bacterial host species, appear to show that jingwei is a dehydrogenase dimer that catalyzes like *Adh* but with altered and diversified substrate binding activity and utilization [66, 67]. This is congruent with other research into the evolution of duplicates, such as within the xanthine dehydrogenase family [68]. One possibility to account for this may be that the gene product folds abnormally and so has lost functional specificity. In any case, as with all chimeric genes, jingwei has retained the core functionality of one or both of its parents but with a reduced pattern of expression.

Retroposed *Adh* mRNA features in two other chimerical genes, *Adh-Twain* and *Adh-Finnegan*, where it has been inserted in different species of *Drosophila*. Interestingly, 230 of the 255 residues contained in the corresponding *Adh* sequences are identical in *Jingwei* and both *Adh-Twain* and *Adh-Finnegan*. Begun and Jones [69] suggest that some sort of convergent adaptation could be at work, but that seems unlikely given that these genes have markedly different patterns of expression [70]—it is perhaps more reasonable to infer that the *Adh* part has undergone the same level of initially relaxed selection followed by reparative compensation. The observed incidence of parallel evolution, as can be seen in Figure 3, something found to be relatively widespread in genetics [71], might be because of a common mutational susceptibility—for which the initial loss of introns associated with the *Adh* part [72] and need for priority readjustments may be a factor. Indeed, research tends to suggest that the presence of introns does have a significant effect on mRNA stability [73]. It is interesting that Begun and Jones infer a burst of evolutionary activity in the early stages but a noticeable slowing down later on. This is consistent with a model of initially relaxed selection in a population increasing in size following a bottleneck. The probability of fixation in a diploid population is $1/2N$ for neutral alleles having no selective (dis)advantage, and so more likely to occur in a smaller set.

3.4. Classical Duplication and Divergence

Perhaps the best example of how duplication and classical evolutionary divergence can facilitate ecological adaptation is the unique case of concerted evolution in colobine monkeys. The animals have adjusted to a predominantly leaf-eating diet by evolving a variant pancreatic ribonuclease (pRNase) recruited to perform a particular role as a digestive enzyme in foregut fermenters [74]. The data suggests that two pRNase paralogs (1A and 1B), both 156 residues long, have been selected for in the colobine monkeys, with one adapting to its role with the loss of positive charge—namely arginine residues. In *colobus polykomos*, the number of acidic residues in this gene product has increased from 13 to 15, whereas those for bases have decreased from 20 to 17. A test for selection revealed evidence in support of a partial gain in function. A total of 15 bp substitutions were identified, 13 of which replaced the ancestral amino acid. However, the number of transitions to transversions was as expected from neutral evolution (11:4), and only five of the residue changes were radical ones.

Therefore, this classical model of gene duplication, mutation, and natural selection would appear to demonstrate how evolutionary processes can

DOI 10.1002/cplx

FIGURE 3

A comparison of retroposed Adh in both Jingewi (Jgw), Adh-Twain (Atw), and Adh-Finnegan (Afg) in different species of Drosophila. The ancestral sequence (Adh) is also given. Adh has evolved in a parallel fashion, where it has been inserted into a host gene by retroposition. A frameshift at the C-term has truncated the protein in Jgw and the others. Clearly, the insertion of the retroduplicated gene has resulted in a very similar pattern of molecular evolution within these separate species. This would suggest a mutational convergence that does involve adaptation other than compensation.

modify and optimize existing information to meet new environmental pressures. However, this also shows how evolutionary divergence is limited and results in closely related and not entirely novel functions. This may also be true for the nuclear receptor family that are comprised of ligand-mediated regulators of gene expression. It is inferred that "molecular tinkering," entailing modifications in ligand specificity due to subtle changes in the the ligand pocket, where the signaling compound binds, led to associations in various duplicate members with other hormones and signals [75].

Another interesting case of classical divergence within a gene family concerns the tetrameric oxygen-binding protein, hemoglobin, found in the red blood cells of vertebrates. Five variants of hemoglobin exist at the β-globin locus cluster in both humans and chimpanzees, all under the control of single regulatory region [76]; each member is differentially expressed throughout the development of the organism: Epislon (*HBE*), for example, is normally expressed only in the embryonic yolk sac. It is precisely for this reason that gene duplication may have been involved in the division and specialization of the original functions of a gene divided among different paralogs—as the organism can not exactly wait for the gene pertinent to the next developmental stage of oxygen metabolism to evolve. The five genes present at the locus (including two *HBG* variants) mare highly similar in sequence and could be the functional equivalents of alternatively spliced isoforms of the original gene.

Indeed, there are reasonable grounds to suppose that gene duplication and mutation may be functionally comparable with the action of alternative splicing in general [77]. For example, in certain species of the genus *Drosophila*, an ancestral sex-biased gene, *JanusA*, uses alternative splicing to encode two slightly different proteins, one present in multiple tissues of both sexes and the other present only in sperm. Duplication of *JanusA* created *JanusB*, which then specialized to encode a sperm-specific protein very similar to the function of the former spliced variant [78]. Therefore, in this situation, no new information was produced.

Subfunctionalization, whereby the information content of a parent gene is differentially partitioned amongst its daughters, is believed to be a common occurrence among surviving duplicates [79]. Here, duplication allows the original functionality of a gene to be spread across more stretches of DNA, although conserving the basic information content contained in the ancestral sequence. Subfunctionalization constitutes a loss in functional redundancy, due to the combination of both complementary degeneration and stabilizing selection, and helps explain why knocking out certain paralogs can have a harmful effect. However, the benefit of this is that a degree of functional specialization can be arrived at which can have gains in efficiency in certain circumstances .In baker's yeast, *Saccharomyces cerevisiae*, two galactose regulatory genes (*GAL1* and *GAL3*) are believed to have evolved from a single bifunctional gene in an ancestral species, resulting in greater flexibility [80].

DOI 10.1002/cplx

3.5. Duplication and Intragenic Amplification: The Case of an AFGP in Notonthenioids

All of the examples above involve evolution within the existing kind as opposed to any divergence that would lead to the emergence of a new type of gene. The first clear attempt at explaining how an old protein gene could spawn a new gene coding for an entirely new protein, and with a distinctly different function, is the case of a trypsinogen to antifreeze glycoprotein (AFGP) conversion in the notothenioid species, *Dissostichus mawsoni* [81]. The ice-binding AFGP that circulates in the blood of the Antarctic fish enables them to avoid freezing in their perpetually icy environment. This crucial survival protein is believed to have evolved from a pancreatic trypsinogen-like protease—a digestive enzyme. Indeed, both proteins are observed to be biosynthesized and secreted in the pancreas [82], and this is reflected in the shared regulatory features found in the UTR and signal peptide. The AFGP is characterized by repeats of two 3-residue components: TAA and TPA. These comprise about 60% of the 362-residue protein, *Dm3l*, one member of the AFGP family. The reasons given for the possible origin of the AFGP from a protease ancestor are:

i. Exon 1 (containing the secretory signal and 5′UTR) in both AFGP and trypsinogen genes is almost identical, as is the 3′ UTR of both genes.
ii. The sequence of intron 1 of the trypsinogen gene is included within as two parts within intron 1 of the AFGP gene.
iii. A 9-nt element in the trypsinogen gene—acagcggca (TAA)—that straddles intron 1 and exon 2 comprises the main repeating unit of the AFGP gene.
iv. The topological proximity of both genes on the same chromosome indicates the likelihood of tandem duplication.
v. The discovery of a chimeric AFGP-protease gene (*Dm7m*) that may be intermediate [83].

Cheng et al. speculate that the ancestral protease gene was converted into the AFGP through a process that involved four major steps: a bulk deletion, intronic (de novo) recruitment, repeated internal amplification, and finally illegitimate recombination. However, this proposed mechanism is unlikely to have occurred for the following reasons:

i. The authors readily acknowledge that the bulk deletion of four exons and four introns is not likely to be tolerated even in a redundant duplicate, as it results in an entirely nonfunctional copy. This would make it liable for complete disintegration by null mutations, and not for its apparently miraculous reincarnation as an entirely new gene.
ii. The AFGP promoter elements at the 5′ flanking sequence upstream of exon 1 are believed to be different from those found in the trypsinogen gene. Both proteins are produced in different amounts and also expressed in a different manner. The proper function and behavior of the glycoprotein depends on changes made or added to the promoter sequences.
iii. Intron 1 in *Dm3l* is 1908 bp long, whereas the corresponding one is 238 bp in the trypsinogen gene. There is no explanation provided for this eightfold difference in size, and the additional sequence's intronic information, other than a huge insertion (e.g., a retrotransposon, for which there is no trace) or a case of repeated intronic amplification.
iv. The authors propose, implausibly, that the repeating TAA and TPA elements—hardly a unique sequence—could have been produced by successive polymerase replication slippage or unequal intragenic recombination dozens of times over. However, this process is both indiscriminate and inefficient [84], and there is no reason to suppose it would selectively and exactly repeat the 9-nt elements, with no resultant frameshift causing a premature termination. The positioning of the proline residues is important as far as protein stability and folding is concerned [85].
v. There is no origin given for the inclusion of the important spacer sequence elements—LIF/LNF/FNF/LNL [86]—other than an unsubstantiated and unfalsifiable claim that they could have been introduced through a yet unspecified "recombinatory event." There is also a nonhomogenous pattern of repetition observed that is not exactly what one would expect from successive amplification.

A key problem associated with the Darwinian mechanism of evolution is that many of the putative incipient and intermediate stages in the development of a biological trait may not be useful themselves and may even be harmful. This is exactly the problem with Cheng's proposed conversion. The incipient stage consists of a bulk deletion that would be almost certainly selected against, despite it being in a gene copy, as the cistron's core information and any useful functional redundancy it may have offered, would have been entirely lost. The resultant protein would be liable to misfold anyway. It is also extremely problematic that the initial intronic recruitment and its subsequent amplification would have been in any way functional—as far as binding to ice crystals or glycosylation is concerned—or have any exaptive utility. The hypothesized metamorphosis would have required widespread and related changes that must have been coordinated and synchronized—and so representing something to the effect of a directional

DOI 10.1002/cplx

saltation. However, this is not something a blind, unsupervised process that can be achieved. It is, however, plausible to suggest that the commonality shared between both genes at their respective termini is indicative of the possibility, at least, that the glycoprotein was derived from an ancestral protease template.

Moreover, the antifreeze proteins that have been found in Arctic cod [87] are completely different in sequence and organization from their Antarctic cousins—this means that the same trypsinogen-like gene could not have been the ancestral gene in this case. Although this is passed off as evidence of "convergent evolution," this serves only to provide another problem as to how a gene believed to be of a more recent origin could have evolved.

3.6. De Novo Recruitment Without Duplication

Although duplication is central to the modern evolutionary synthesis, in recent years, the possibility that previously extragenic, noncoding regions of DNA could be recruited wholesale to become translated as functioning proteins, as opposed to just minor exonization observed in the formation of the amino end of the *Sdic* gene. This represents a return to the idea of the hopeful monster [88] at the molecular level. For example, such origination has been proposed in the case of the yeast gene *BSC4* [89] (of unknown function); and the human upregulated gene *CLLU1* [90] that is believed to have some role in pathogenesis of chronic lymphomatic leukemia and shares structural motifs with the cytokine, *IL-4*, that is used in the immune system [91]. In the case of *CLLU1*, a single nucleotide deletion of adenine in a stretch of DNA orthologous with chimpanzees has created a frameshift and expanded ORF, large enough to be fully functional when translated as a protein. However, this inference may be incorrect. Rather than the deletion creating a new stretch of translated DNA, it is likely that a back mutation restored the original ORF that became essentially divided in two as a result of an insertion—a very common phenomenon observed in indel-induced frameshifts [92]. Thus, far from being a case of bulk de novo recruitment of ncDNA, *CLLU1* in humans is a gene that may have been fully reactivated while still inactive in other primate lineages. The corresponding gene in chimpanzees, if transcribed and regulated, may still be partially functional as two potential 42-codon reading frames are preserved at either terminus. Thus, the de novo and fortuitous origination of entire reading frames may be a profound misinterpretation of cases of pseudogenes being reactivated.

Alternatively, functional sections of noncoding DNA, or perhaps even "dormant" reading frames, have become translated into proteins that perform a particular task. There is indeed evidence for the existence of ORFs within introns [93] and other regions of noncoding DNA [94] that may be the result of transposition events. However, another possibility is that instead "junk" sequences of ncDNA are accidentally transcribed and translated into nonfunctional products that are fixed by neutral evolution, and which serve no purpose, other than perhaps being assigned to the cell's garbage collection and recycling system. In any case, as a mechanism for the creation of novel motifs and protein domains, de novo recruitment of noncoding DNA would seem extremely improbable and implausible.

4. CONCLUSION

Gene duplication and subsequent evolutionary divergence certainly adds to the size of the genome and in large measure to its diversity and versatility. However, in all of the examples given above, known evolutionary mechanisms were markedly constrained in their ability to innovate and to create any novel information. This natural limit to biological change can be attributed mostly to the power of purifying selection, which, despite being relaxed in duplicates, is nonetheless ever-present. The reason for this stabilization of function is not obvious, although the role of duplicates in compensating for deleterious loss of function mutation at paralogous sites may be an important factor. Likewise, there exists a preservation of ancestral functions through the process of a differential division of labor among duplicates, namely that of subfunctionalization. Moreover, both the possibility and opportunity for beneficial changes leading to major functional innovations was found to be not especially convincing. For example, duplicate enzyme-coding genes tend to retain the same ancestral catalytic activity and simply apply that function to different substrates, often by partial degradation of function and the loss of the precise specificity of the parent. However, these may prove to have an important adaptive value in response to environmental challenges such as with respect to temperature, drought, pathogens, and UV radiation.

Where substantive sequence evolution had occurred, it could have been because a respite in selective constraints led to significant degeneration. In the case of *Sdic* and *Jingwei*, both genes evolved from duplicates affected by significant deletions or the silencing of exonic information and were then co-opted for use in a different context. This development has likely been misinterpreted in many cases as evidence of a gain in information under positive Darwinian selection, especially when extensive compensatory changes are involved that can amplify sequence divergence in the process. In this sense, a proclivity toward functional stability and the conservation of information, as opposed to any adventurous innovation, predominates.

The various postduplication mechanisms entailing random mutations and

DOI 10.1002/cplx

recombinations considered were observed to tweak, tinker, copy, cut, divide, and shuffle existing genetic information around, but fell short of generating genuinely distinct and entirely novel functionality. Contrary to Darwin's view of the plasticity of biological features, successive modification and selection in genes does indeed appear to have real and inherent limits: it can serve to alter the sequence, size, and function of a gene to an extent, but this almost always amounts to a variation on the same theme—as with *RNASE1B* in colobine monkeys. The conservation of all-important motifs within gene families, such as the homeobox or the MADS-box motif, attests to the fact that gene duplication results in the copying and preservation of biological information, and not its transformation as something original.

The case of evolution in notothenioid fish, entailing the speculative conversion of a protease duplicate into an AFGP, only serves to demonstrate the huge problem of supposing that cumulative random changes would contrive to produce novel information, especially if major deletions and other degenerative mutations were involved.

Although the focus here has been on the information within exons that code for the amino acid sequences in proteins, noncoding DNA—which comprises the vast majority of the molecule—also contains information necessary for the regulation and expression of gene products. Changes in these regions can have a profound effect on an organism's evolution. But, although important, without a repertoire of proteins with which to regulate, this is ancillary in effect. For example, it is impossible for an organism to develop vision without the exons coding for light-sensitive opsins or feathers for flight without the presence of keratins in the skin.

Gradual natural selection is no doubt important in biological adaptation and for ensuring the robustness of the genome in the face of constantly changing environmental pressures. However, its potential for innovation is greatly inadequate as far as explaining the origination of the distinct exonic sequences that contribute to the complexity of the organism and diversity of life. Any alternative/revision to Neo-Darwinism [95] has to consider the holistic nature and organization of information encoded in genes, which specify the interdependent and complex biochemical motifs that allow protein molecules to fold properly and function effectively.

REFERENCES

1. Glansdorff, N.; Xu, Y.; Labedan, B. The last universal common ancestor: Emergence, constitution and genetic legacy of an elusive forerunner. Biol Direct 2008, 3, 29.
2. Salthe, S.N. Natural selection in relation to complexity. Artif Life 2008, 14, 363–374.
3. Silander, O.K.; Ackermann, M. The constancy of gene conservation across divergent bacterial orders. BMC Res Notes 2009, 2, 2.
4. Murphy, R.; Tsai, A., Eds. Protein Folding, Misfolding, Stability, and Aggregation. Biotechnology Progress 2007, 23, 548–552.
5. Sawyer, S.A.; Parsch, J.; Zhang, Z.; Hartl, D.L. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci USA 2007, 104, 6504–6510.
6. Burke, M.K.; Dunham, J.P.; Shahrestani, P.; Thornton, K.R.; Rose, M.R.; Long, A.D. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature 2010, 467, 587–590.
7. Sarah, O.P. Two steps forward, one step back: The pleiotropic effects of favoured alleles. Proc Biol Sci 2004, 271, 705–714.
8. Nozawa, M.; Suzuki, Y. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA 2009, 106, 6700–6705.
9. Galtier, N.; Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 2007, 23, 273–277.
10. Betancourt, A.J.; Presgraves, D.C. Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci USA 2002, 99, 13616–13620.
11. Pagani, F.; Raponi, M.; Baralle, F.E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA 2005, 102, 6368–6372.
12. Grange, T.; de Sa, C.M.; Oddos, J.; Pictet, R. Human mRNA polyadenylate binding protein: Evolutionary conservation of a nucleic acid binding motif. Nucleic Acids Res 1987, 15, 4771–4787.
13. Jordan, I.K.; Wolf, Y.I.; Koonin, E.V. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol 2004, 4, 22.
14. Zhang, J.; Rosenberg, H.F.; Nei, M. Proc Natl Acad Sci USA 1998, 95, 3708–3713.
15. Taverna, D.M.; Goldstein, R.M. The evolution of duplicated genes considering protein stability constraints. Pac Symp Biocomput 2000, 69–80.
16. Regis, S.; Grossi, S.; Corsolini, F.; Biancheri, R.; Filocamo, M. PLP1 gene duplication causes overexpression and alteration of the PLP/DM20 splicing balance in fibroblasts from Pelizaeus-Merzbacher disease patients. Biochim Biophys Acta 2009, 1792, 548–554.
17. Edger, P.P.; Pires, J.C. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res 2009, 17, 699–717.
18. Teichmann, S.A.; Babu, M.M. Gene regulatory network growth by duplication. Nat Genetics 2004, 36, 492–496.

DOI 10.1002/cplx

19. Chain, F.J.J.; Ilieva, D.; Evans, B.J. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. BMC Evol Biol 2008, 8, 43.
20. Hanada, K.; Kuromori, T.; Myouga, F.; Toyoda, T.; Li, W.H.; Shinozaki, K. Evolutionary persistence of functional compensation by duplicate genes in arabidopsis. Genome Biol Evol 2009, 1, 409–414.
21. Hsiao, T.L.; Vitkup, D. Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 2008, 4, e1000014.
22. Gu, Z.; Steinmetz, L.M.; Gu, X.; Scharfe, C.; Davis, R.W.; Li, W.H. Role of duplicate genes in genetic robustness against null mutations. Nature 2003, 421, 63–66.
23. Vavouri, T.; Semple, J.I.; Lehner, B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet 2008, 24, 485–488.
24. Qian, W.; Liao, B.Y.; Chang, A.Y.; Zhang, J. Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet 2010, 26, 425–430.
25. MacCarthy, T.; Bergman, A. The limits of subfunctionalization. BMC Evol Biol 2007, 7, 213.
26. Kimura, M. On the probability of fixation of mutant genes in a population. Genetics 1962, 47, 713–719.
27. Hood, L.; Galas, D. The digital code of DNA. Nature 2003, 421, 444–448.
28. Godfrey-Smith, P. Information in Biology. The Cambridge Companion to the Philosophy of Biology. Camrbridge: Cambridge University Press, 2007, pp. 103–111.
29. Shannon, C.E. A mathematical theory of communication. Bell Syst Tech J 1948, 27, 379–423, 623–656.
30. Kacprzak, M.M.; Than, M.E.; Juliano, L.; Juliano, M.A.; Bode, W.; Lindberg, I. Mutations of the PC2 substrate binding pocket alter enzyme specificity. J Biol Chem 2005, 280, 31850–31858.
31. Newcomb, R.D.; Campbell, P.M.; Ollis, D.L.; Cheah, E.; Russell, R.J.; Oakeshott, J.G. A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. Proc Natl Acad Sci USA 1997, 94, 7464–7468.
32. Resch, A.M.; Carmel, L.; Mariño-Ramírez, L.; Ogurtsov, A.Y.; Shabalina, S.A.; Rogozin, I.B.; Koonin, E.V. Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol 2007, 24, 1821–1831.
33. Sangar, V.; Blankenberg, D.J.; Altman, N.; Lesk, A.M. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 2007, 8, 294.
34. Fay, J.C.; Wu, C.I. Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genomics Hum Genet 2003, 4, 213–235.
35. Delcour, A.H. Outer membrane permeability and antibiotic resistance. Biochim Biophys Acta 2009, 1794, 808–816.
36. Zeng, K.; Fu, Y.X.; Shi, S.; Wu, C.I. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 2006, 174, 1431–1439.
37. Charlesworth, B.; Morgan, M.T.; Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993, 134, 1289–1303.
38. Mekel-Bobrov, N. The ongoing adaptive evolution of ASPM and Microcephalin is not explained by increased intelligence. Hum Mol Genet 2007, 16, 600–608.
39. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998, 15, 568–573.
40. Huelsenbeck, J.P.; Jain, S.; Frost, S.W.; Pond, S.L.A. Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci USA 2006, 103, 6263–6268.
41. Yang, Z.; Bielawski, J.P. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000, 15, 496–503.
42. Davis, B.H.; Poon, A.F.; Whitlock, M.C. Compensatory mutations are repeatable and clustered within proteins. Proc Biol Sci 2009, 276, 1823–1827.
43. Dagan, T.; Talmor, Y.; Graur, D. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol Biol Evol 2002, 19, 1022–1025.
44. Cordaux, R.; Batzer, M.A. The impact of retrotransposons on human genome evolution. Nat Rev Genet 2009, 10, 691–703.
45. Ponce, R.; Hartl, D.L. The evolution of the novel Sdic gene cluster in *Drosophila melanogaster*. Genetica 2006, 376, 174–183.
46. Ranz, J.M.; Ponce, A.R.; Hartl, D.L.; Nurminsky, D. Origin and evolution of a new gene expressed in the *Drosophila* sperm axoneme. Genetica 2003, 188, 233–244.
47. Ranz, J.M.; Ponce, A.R.; Hartl, D.L.; Nurminsky, D. Origin and evolution of a new gene expressed in the *Drosophila* sperm axoneme Genetica 2003, 118, 233–244.
48. Nurminsky, D.I.; Nurminskaya, M.V.; De Aguiar, D.; Hartl, D.L. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 1998, 396, 572–575.
49. Nurminsky, D.I.; Hartl, D.L. Reply: How was the *Sdic* gene fixed? Nature 1999, 400, 519–520.
50. Ponce, R. The recent origin of the *Sdic* gene cluster in the *melanogaster* subgroup. Genetica 2009, 135, 415–418.
51. Nei, M. Bottlenecks, genetic polymorphism and speciation. Genetics 2005, 170, 1–4.
52. Paulding, C.A.; Ruvolo, M.; Haber, D.A. The Tre2 (USP6) oncogene is a hominoid-specific gene. Proc Natl Acad Sci USA 2003, 100, 2507–2511.
53. Long, M. A new function evolved from gene fusion. Genome Res 2000, 10, 1655–1657.
54. Ohno, S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci USA 1984, 81, 2421–2425.

DOI 10.1002/cplx

55. Negoro, S.; Ohki, T.; Shibata, N.; Mizuno, N.; Wakitani, Y.; Tsurukame, J.; Matsumoto, K.; Kawamoto, I.; Takeo, M.; Higuchi, Y. X-ray crystallographic analysis of 6-aminohexanoate-dimer hydrolase: Molecular basis for the birth of a nylon oligomer-degrading enzyme. J Biol Chem 2005, 280, 39644–39652.
56. Kinoshita, S. Purification and characterization of 6-aminohexanoic-acid-oligomer hydrolase of Flavobacterium sp. Ki72. Eur J Biochem 1981, 116, 547–551.
57. Okamura, K.; Feuk, L.; Marquès-Bonet, T.; Navarro, A.; Scherer, S.W. Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics 2006, 88, 690–697.
58. Shiba, K.; Takahashi, Y.; Noda, T. Creation of libraries with long ORFs by polymerization of a microgene. Proc Natl Acad Sci USA 1997, 94, 3805–3810.
59. Horne, M.C.; Goolsby, G.L.; Donaldson, K.L.; Tran, D.; Neubauer, M.; Wahl, A.F. Cyclin G1 and cyclin G2 comprise a new family of cyclins with contrasting tissue-specific and cell cycle-regulated expression. J Biol Chem 1996, 271, 6050–6061.
60. Javaud, C.; Dupuy, F. The fucosyltransferase gene family: An amazing summary of the underlying mechanisms of gene evolution. Genetica 2003, 118, 157–170.
61. Long, M.; Langle, C.H. Natural selection and the origin of Jingwei, a chimeric processed functional gene in *Drosophila*. Science 1993, 260, 91–95.
62. Wang, W.; Zhang, J.; Alvarez, C.; Llopart, A.; Long, M. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. Mol Biol Evol 2000, 17, 1294–1301.
63. Betrán, E. Gene Fusion. Encyclopedia of Life Sciences; John Wiley & Sons, Ltd: Chichester, 2008.
64. Bruce, R.; Levin, E.; Véronique, P.; Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. Genetics 2000, 154, 985–997.
65. Björkman, J.; Nagaev, I.; Berg, O.G.; Hughes, D.; Andersson, D.I. Effects of environment on compensatory mutations to ameliorate costs of antibiotic resistance. Science 2000, 287, 1479–1482.
66. Zhang, J.; Dean, A.M.; Brunet, F.; Long, M. Evolving protein functional diversity in new genes of *Drosophila*. Proc Natl Acad Sci USA 2004, 101, 16246–16250.
67. Zhang, J.; Yang, H.; Long, M.; Li, L.; Dean, A.M. Evolution of enzymatic activities of testis-specific short-chain dehydrogenase/reductase in *Drosophila*. J Mol Evol 2010, 71, 241–249.
68. Rodríguez-Trelles, F.; Tarrío, R.; Ayala, F.J. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc Natl Acad Sci USA 2003, 100, 13413–13417.
69. Jones, C.D.; Begun, D.J. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci USA 2005, 102, 11373–11378.
70. Jones, C.D.; Custer, A.W.; Begun, D.J. Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics 2005, 170, 207–219.
71. Rokas, A.; Carroll, S.B. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 2008, 25, 1943–1953.
72. Llopart, A.; Comeron, J.M.; Brunet, F.G.; Lachaise, D.; Long, M. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc Natl Acad Sci USA 2002, 99, 8121–8126.
73. Wang, H.F.; Feng, L.; Niu, D.K. Relationship between mRNA stability and intron presence. Biochem Biophys Res Commun 2007, 354, 203–208.
74. Zhang, J. Parallel functional changes in the digestive RNAses of ruminants and colobines by divergent amino acid substitutions. Mol Biol Evol 2003, 20, 1310–1317.
75. Thornton, J.W.; Bridgham, J.T.; Eick, G.N.; Larroux, C.; Deshpande, K.; Harms, M.J.; Gauthier, M.E.; Ortlund, E.A.; Degnan, B.M.; Thornton, J.W. Protein evolution by molecular tinkering: Diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. PLoS Biol 2010, 8, e1000497.
76. Caterina, J.J. Multiple elements in human beta-globin locus control region 5′ HS 2 are involved in enhancer activity and position-independent, transgene expression. 1994, 22, 1006–1011.
77. Talavera, D.; Vogel, C.; Orozco, M.; Teichmann, S.A.; de la Cruz, X. The (in) dependence of alternative splicing and gene duplication. PloS Comput Biol 2007, 3, e33.
78. Parsch, J.; Meiklejohn, C.D.; Hauschteck-Jungen, E.; Hunziker, P.; Hartl, D.L. Molecular evolution of the ocnus and janus genes in the *Drosophila melanogaster* species subgroup. Mol Biol Evol 2001, 18, 801–811.
79. Lynch, M.; Conery, J.S. The evolutionary fate and consequences of duplicate genes. Science 2000, 290, 1151–1155.
80. Hittinger, C.T.; Carroll, S.B. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 2007, 449, 677–681.
81. Chen, L.; DeVries, A.L.; Cheng, C.C.-H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci USA 1997, 94, 3811–3816.
82. Cheng, C.-H.C.; Cziko, P.A.; Evans, C.W. Non-hepatic origin of notothenioid antifreeze reveals pancreatic synthesis as common mechanism in polar fish freezing avoidance. Proc Natl Acad Sci USA 2006, 103, 10491–10496.
83. Chen, H.C.; Chen, L. Evolution of an antifreeze glycoprotein. Nature 1999, 401, 443–444.
84. Viguera, E.; Canceill, D.; Ehrlich, S.D. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J 2001, 20, 2587–2595.
85. Samuel, D.; Kumar, T.K.; Ganesh, G.; Jayaraman, G.; Yang, P.W.; Chang, M.M.; Trivedi, V.D.; Wang, S.L.; Hwang, K.C.; Chang, D.K.; Yu, C. Proline inhibits aggregation during protein refolding. Protein Sci 2000, 9, 344–352.
86. Hsiao, K.; Cheng, C.; Fernandes, I.E.; Detrich, H.W.; DeVries, A.L. An antifreeze glycopeptide gene from the antarctic cod *Notothenia coriiceps* neglecta encodes a polyprotein of high peptide copy number. Proc Natl Acad Sci USA 1990, 88, 2966.

DOI 10.1002/cplx

87. Chen, L.; DeVries, A.L.; Cheng, C.H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc Natl Acad Sci USA 1997, 94, 3817–3822.
88. Theissen, G. The proper place of hopeful monsters in evolutionary biology. Theory Biosci 2006, 124, 349–369.
89. Cai, J.; Zhao, R.; Jiang, H.; Wang, W. De novo origination of a new protein-coding gene in *Sacchoramyces cerevisiae*. Genetics 2008, 179, 487–496.
90. Knowles, D.; McLysaght, A. Recent de novo origin of human protein-coding genes. Genome Res 2009, 19, 1752–1759.
91. Caligaris-Cappio, F. A novel gene for an old disease. Blood 2006, 107, 2594.
92. Nishikawa, T.; Murakami, M.; Hayashi, Ki.; Sato, H.; Otsuki, T.; Kasahara, N.; Yasuda, T.; Kimura, K.; Nagai, K.; Irie, R.; Sugiyama, T.; Isogai, T.; Dunker, A.K.; Konagaya, A.; Miyano, S.; Takagi, T. An extracting system of accurate ORFs from cDNA sequences. Genome Inform 2002, 13, 545–547.
93. Cioffi, A.V.; Ferrara, D.; Cubellis, M.V.; Aniello, F.; Corrado, M.; Liguori, F.; Amoroso, A.; Fucci, L.; Branno, M. An open reading frame in intron seven of the sea urchin DNA-methyltransferase gene codes for a functional AP1 endonuclease. Biochem J 2002, 365, 833–840.
94. Mackiewicz, P.; Kowalczuk, M.; Gierlik, A.; Dudek, M.R.; Cebrat, S. Origin and properties of non-coding ORFs in the yeast genome. Nucleic Acids Res 1999, 27, 3503–3509.
95. Batten, D.; Salthe, S.; Boschetti, F. Visions of evolution: Self-organization proposes what natural selection disposes. Biol Theory 2009, 3, 17–29.

DOI 10.1002/cplx